Characterizing the effect of genetic variants within promoters with distal enhancer functions

Alejandra Medina Rivera, PhD
Laboratorio Internacional de Investigación en Genoma Humano
Universidad Nacional Autónoma de México
November 26th, 2019
Barcelona
Thanks
LIIGH

- International Laboratory for Human Genome Research
  - Founded in 2015 with the aim to attract young researchers.
  - Regulatory Genomics and Bioinformatics Laboratory

http://www.liigh.unam.mx/amedina
@AleMedinaRivera amedina@liigh.unam.mx
Regulatory Genome

Histones
Regulatory Genome

Histones

Transcriptional Factors
Regulatory Genome

Histones

Transcriptional Factors
Annotating the Regulatory Genome

Modified from Huayun Hou
Published Genome-Wide Associations since 12/2012
Published GWA at $p \leq 5 \times 10^{-8}$ for 17 trait categories

- More than 17,000 reported associated SNPs
- 93% in non-coding sequence
Published Genome-Wide Associations at p≤5×10^{-8} for 17 trait categories 11/2019

- More than 159,000 SNPs
- More than 90% fall within non-coding sequence

NIH National Human Genome Research Institute
EMBL-EBI
NHGRI-EBI GWAS Catalog
www.ebi.ac.uk/gwas
Annotating the Regulatory Genome

Modified from Huayun Hou
First characterized Enhancers

<table>
<thead>
<tr>
<th>Gene</th>
<th>Origin</th>
<th>Size (bp)</th>
<th>Distance from TSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Early gene</td>
<td>SV40</td>
<td>196</td>
<td>~200</td>
</tr>
<tr>
<td>Early gene</td>
<td>Cytomegalovirus</td>
<td>406</td>
<td>−524 to −118</td>
</tr>
<tr>
<td>Hsp70</td>
<td>Xenopus</td>
<td>160</td>
<td>−260 to −100</td>
</tr>
<tr>
<td>Fos</td>
<td>Human</td>
<td>340</td>
<td>−404 to −64</td>
</tr>
<tr>
<td>hMT-IIA</td>
<td>Human</td>
<td>327</td>
<td>−366 to −39</td>
</tr>
<tr>
<td>Mmt-IA</td>
<td>Mouse</td>
<td>114</td>
<td>−187 to −73</td>
</tr>
<tr>
<td></td>
<td></td>
<td>155</td>
<td>−194 to −39</td>
</tr>
<tr>
<td>H2A</td>
<td>Urchin</td>
<td>28</td>
<td>−139 to −111</td>
</tr>
<tr>
<td>IFNb</td>
<td>Human</td>
<td>40</td>
<td>−77 to −37</td>
</tr>
</tbody>
</table>

Boshart, M. et al. (1985)  
Deschamps, J. et al. (1985)  
Serfling, E. et al. (1985)  
Serfling, E. et al. (1985)  
Goodbourn, S. et al. (1985)
### Regulatory Regions: Enhancers vs Promoters

<table>
<thead>
<tr>
<th>Gene</th>
<th>Origin</th>
<th>Size (bp)</th>
<th>Distance from TSS</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Early gene</td>
<td>Cytomegalovirus</td>
<td>406</td>
<td>−524 to −118</td>
<td>Boshart, M. et al. (1985)</td>
</tr>
<tr>
<td>Fos</td>
<td>Human</td>
<td>340</td>
<td>−404 to −64</td>
<td>Deschamps, J. et al. (1985)</td>
</tr>
<tr>
<td>hMT-IIA</td>
<td>Human</td>
<td>327</td>
<td>−366 to −39</td>
<td>Serfling, E. et al. (1985)</td>
</tr>
<tr>
<td>Mmt-IA</td>
<td>Mouse</td>
<td>114</td>
<td>−187 to −73</td>
<td>Serfling, E. et al. (1985)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>155</td>
<td>−194 to −39</td>
<td></td>
</tr>
<tr>
<td>IFNb</td>
<td>Human</td>
<td>40</td>
<td>−77 to −37</td>
<td>Goodbourn, S. et al. (1985)</td>
</tr>
</tbody>
</table>
### How to classify Regulatory Regions?

<table>
<thead>
<tr>
<th>Features (active elements)</th>
<th>Promoter</th>
<th>Enhancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intrinsic property</td>
<td>Induce transcription of a heterologous reporter gene</td>
<td>Activate a distal (heterologous) promoter</td>
</tr>
</tbody>
</table>
How to classify Regulatory Regions?

<table>
<thead>
<tr>
<th>Features (active elements)</th>
<th>Promoter</th>
<th>Enhancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intrinsic property</td>
<td>Induce transcription of a heterologous reporter gene</td>
<td>Activate a distal (heterologous) promoter</td>
</tr>
<tr>
<td>Transcription initiation</td>
<td>Unidirectional or divergent</td>
<td>Mainly divergent</td>
</tr>
</tbody>
</table>
## How to classify Regulatory Regions?

<table>
<thead>
<tr>
<th>Features (active elements)</th>
<th>Promoter</th>
<th>Enhancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intrinsic property</td>
<td>Induce transcription of a heterologous reporter gene</td>
<td>Activate a distal (heterologous) promoter</td>
</tr>
<tr>
<td>Transcription initiation</td>
<td>Unidirectional or divergent</td>
<td>Mainly divergent</td>
</tr>
<tr>
<td>Ratio between sense and antisense transcripts</td>
<td>Biased toward sense transcription</td>
<td>Equilibrated</td>
</tr>
</tbody>
</table>

Medina-Rivera, *Trends in Biochemical Sciences*, 2018
## How to classify Regulatory Regions?

<table>
<thead>
<tr>
<th>Features (active elements)</th>
<th>Promoter</th>
<th>Enhancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intrinsic property</td>
<td>Induce transcription of a heterologous reporter gene</td>
<td>Activate a distal (heterologous) promoter</td>
</tr>
<tr>
<td>Transcription initiation</td>
<td>Unidirectional or divergent</td>
<td>Mainly divergent</td>
</tr>
<tr>
<td>Ratio between sense and antisense transcripts</td>
<td>Biased toward sense transcription</td>
<td>Equilibrated</td>
</tr>
<tr>
<td>Transcription elongation</td>
<td>Produce long polyadenylated transcripts</td>
<td>Some enhancers can produce low levels of polyadenylated transcripts</td>
</tr>
</tbody>
</table>

Medina-Rivera, *Trends in Biochemical Sciences*, 2018
## How to classify Regulatory Regions?

<table>
<thead>
<tr>
<th>Features (active elements)</th>
<th>Promoter</th>
<th>Enhancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intrinsic property</td>
<td>Induce transcription of a heterologous reporter gene</td>
<td>Activate a distal (heterologous) promoter</td>
</tr>
<tr>
<td>Transcription initiation</td>
<td>Unidirectional or divergent</td>
<td>Mainly divergent</td>
</tr>
<tr>
<td>Ratio between sense and antisense transcripts</td>
<td>Biased toward sense transcription</td>
<td>Equilibrated</td>
</tr>
<tr>
<td>Transcription elongation</td>
<td>Produce long polyadenylated transcripts</td>
<td>Some enhancers can produce low levels of polyadenylated transcripts</td>
</tr>
<tr>
<td>Histone modifications</td>
<td>H3K27ac (H3K4me1 &lt; H3K4me3)</td>
<td>H3K27ac (H3K4me1 &gt; H3K4me3)</td>
</tr>
</tbody>
</table>
### How to classify Regulatory Regions?

<table>
<thead>
<tr>
<th>Features (active elements)</th>
<th>Promoter</th>
<th>Enhancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intrinsic property</td>
<td>Induce transcription of a heterologous reporter gene</td>
<td>Activate a distal (heterologous) promoter</td>
</tr>
<tr>
<td>Transcription initiation</td>
<td>Unidirectional or divergent</td>
<td>Mainly divergent</td>
</tr>
<tr>
<td>Ratio between sense and antisense transcripts</td>
<td>Biased toward sense transcription</td>
<td>Equilibrated</td>
</tr>
<tr>
<td>Transcription elongation</td>
<td>Produce long polyadenylated transcripts</td>
<td>Some enhancers can produce low levels of polyadenylated transcripts</td>
</tr>
<tr>
<td>Histone modifications</td>
<td>H3K27ac (H3K4me1 &lt; H3K4me3)</td>
<td>H3K27ac (H3K4me1 &gt; H3K4me3)</td>
</tr>
<tr>
<td>RNAPII and GTF</td>
<td>Present</td>
<td>Present</td>
</tr>
</tbody>
</table>
### How to classify Regulatory Regions?

<table>
<thead>
<tr>
<th>Features (active elements)</th>
<th>Promoter</th>
<th>Enhancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intrinsic property</td>
<td>Induce transcription of a heterologous reporter gene</td>
<td>Activate a distal (heterologous) promoter</td>
</tr>
<tr>
<td>Transcription initiation</td>
<td>Unidirectional or divergent</td>
<td>Mainly divergent</td>
</tr>
<tr>
<td>Ratio between sense and antisense transcripts</td>
<td>Biased toward sense transcription</td>
<td>Equilibrated</td>
</tr>
<tr>
<td>Transcription elongation</td>
<td>Produce long polyadenylated transcripts</td>
<td>Some enhancers can produce low levels of polyadenylated transcripts</td>
</tr>
<tr>
<td>Histone modifications</td>
<td>H3K27ac (H3K4me1 &lt; H3K4me3)</td>
<td>H3K27ac (H3K4me1 &gt; H3K4me3)</td>
</tr>
<tr>
<td>RNAPII and GTF</td>
<td>Present</td>
<td>Present</td>
</tr>
<tr>
<td>CpG islands</td>
<td>Majority</td>
<td>Very rare</td>
</tr>
</tbody>
</table>
How to assess the enhancer function of promoters?
Major challenge: How to assess the enhancer function of promoters?
Detection of long-range regulatory regions
Enhancers

Figure 12-21
Introduction to Genetic Analysis, Eleventh Edition
© 2015 W. H. Freeman and Company
Detection of long-range regulatory regions
Enhancers

- **Starr-seq**: High-throughput assessment of sequence enhancer potential

Arnold, C.D. et al. (2013). Science (80- ).,

Detection of long-range regulatory regions

Enhancers

- Starr-seq: High-throughput assessment of sequence enhancer potential

Arnold, C.D. et al. (2013). Science (80-).

Detection of long-range regulatory regions
Enhancers

- Starr-seq: High-throughput assessment of sequence enhancer potential

Arnold, C.D. et al. (2013). Science (80-).

Detection of long-range regulatory regions
Enhancers

- Starr-seq: High-throughput assessment of sequence enhancer potential

Arnold, C.D. et al. (2013). Science (80-).

Major challenge: How to assess the enhancer function of promoters?
Major challenge: How to assess the enhancer function of promoters?
Can promoters work as enhancers?

CapStarr-seq

Salvatore Spicuglia

RefSeq promoters

↓ Capture

↓ Cloning

GFP

pA

↓ Transfection

HeLa  K562

Targeted RNA-seq

Lan Dao, et al, Nature Genetics. 2017
Can promoters work as enhancers?

Lan Dao, et al, Nature Genetics. 2017
How many promoters work also as enhancers?

Total of ePromoters:
- K562 = 632 (3%)
- HELA = 493 (2.37%)
- Total analyzed promoters = 20,719 (100%)
RNA expression of genes downstream promoters

Lan Dao, et al, Nature Genetics. 2017
RNA expression of genes downstream promoters

Lan Dao, et al, Nature Genetics. 2017
Epromoters had a higher H3K27ac/H3K4me3 ratio in the cell type where they were found to be active

Lan Dao, et al, Nature Genetics. 2017
Epromoters had a higher H3K27ac/H3K4me3 ratio in the cell type where they were found to be active.
Functional annotation of genes downstream Epromoters
Epromoters after interferon stimuli
Epromoters after interferon stimuli

Genomic library

RefSeq promoters

Capture

Cloning

GFP

pA

Transfection

Interferon

K562

Targeted RNA-seq
Epromoters after interferon stimuli
Which are the regulatory mechanisms affecting Epromoters?
Transcription factors control gene regulation by binding to specific DNA sequences.

- Transcription Factors interact with DNA binding to sequence specific sites.
Transcription Factors enriched motifs in Epromoters

K562

- Epromoters
- non-Epromoters

**position-scan**: Motif enrichment in a given position of a set of sequences

Lan Dao, et al, Nature Genetics. 2017
Transcription Factors enriched motifs in Epromoters
Which are the genetic variants that could be affecting Epromoters?
• Variants associated with human traits were downloaded from the Genome Wide Association Studies (GWAS) database.

• Additionally for GWAS SNPs variants that were in LD with an $r^2$ of 0.8 and within a 1 Mb window were included in the study.
Variants associated with human traits were downloaded from the Genome Wide Association Studies (GWAS) database.

Additionally for GWAS SNPs variants that were in LD with an $r^2$ of 0.8 and within a 1 Mb window were included in the study.

We found 22 GWAS SNPs or SNPs LD within ePromoters.
Human Genetic Variants associated to traits annotations

SNP SNP GWAS Gut SNP SNP

SNP SNP GWAS Hematological measurements SNP SNP

ePromoter ePromoter
Human Genetic Variants associated to traits annotations
### Hematological measurements

<table>
<thead>
<tr>
<th></th>
<th>HELA</th>
<th>Label</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active</td>
<td>4</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>Inactive</td>
<td>420</td>
<td>21646</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>K562</th>
<th>Label</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active</td>
<td>10</td>
<td>345</td>
<td></td>
</tr>
<tr>
<td>Inactive</td>
<td>628</td>
<td>32503</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>OddsRatio</th>
<th>Pvalue,adj</th>
</tr>
</thead>
<tbody>
<tr>
<td>HELA</td>
<td>15.853</td>
</tr>
<tr>
<td>K562</td>
<td>1.5001</td>
</tr>
</tbody>
</table>
### Other measurements

<table>
<thead>
<tr>
<th>HELA</th>
<th>Label</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active</td>
<td>2</td>
<td>15</td>
</tr>
<tr>
<td>Inactive</td>
<td>2021</td>
<td>1345</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>K562</th>
<th>Label</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active</td>
<td>299</td>
<td>55</td>
</tr>
<tr>
<td>Inactive</td>
<td>30426</td>
<td>2705</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>HELA</th>
<th>Odds Ratio</th>
<th>P-value adj</th>
</tr>
</thead>
<tbody>
<tr>
<td>HELA</td>
<td>0.012146</td>
<td>3.01e-13</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>K562</th>
<th>Odds Ratio</th>
<th>P-value adj</th>
</tr>
</thead>
<tbody>
<tr>
<td>K562</td>
<td>0.4747</td>
<td>7.58e-05</td>
</tr>
</tbody>
</table>
What is the effect of GWAS variants that disrupt TF binding?

- Genetic variants in gene regulatory regions are known to cause many diseases
- Can be informative of transcription factors and regulatory mechanisms
What is the effect of GWAS variants that disrupt TF binding?

- Genetic variants in gene regulatory regions are known to cause many diseases
- Can be informative of transcription factors and regulatory mechanisms
Detecting GWAS variants effect on TF binding

- variation-scan: PSSM based algorithm as part of the RSAT suite
- Flexible tool that can be used to scan any variable with any motif.
- Multiple organisms are available.

Weight Score

**REF:** 8.1  
**MUT:** 2.1  
**DIF:** 6.0

Walter Santana
Detecting protein disruption of binding at base pair resolution

A) Genome-wide variants

B) Variation sequence reconstruction

C) Assess impact

Variation-tools

User input

RSAT available data sets
Confirmatory Results

**RSAT Var-tools: an accessible and flexible framework to predict the impact of regulatory variants on transcription factor binding**

Walter Santana-Garcia, Maria Rocha-Acevedo, Lucia Ramirez-Navarro, Yvon Mbouamboua, Denis Thieffry, Morgane Thomas-Chollier, Bruno Contreras-Moreira, Jacques van Helden, Alejandra Medina-Rivera

doi: https://doi.org/10.1101/623090

This article is a preprint and has not been certified by peer review [what does this mean?].
Detecting GWAS variants effect on TF binding

Confirmatory Results

RSAT Var-tools: an accessible and flexible framework to predict the impact of regulatory variants on transcription factor binding

Walter Santana-Garcia, Maria Rocha-Acevedo, Lucia Ramirez-Navarro, Yvon Mbouamboua, Denis Thieffry, Morgane Thomas-Chollier, Bruno Contreras-Moreira, Jacques van Helden, Alejandra Medina-Rivera

doi: https://doi.org/10.1101/623090

This article is a preprint and has not been certified

Abstract  Full Text  Info/History
GWAS and LD-GWAS SNPs affect TF binding

- 109 TF binding - SNP in HELA ePrompters
- 190 in K562
- 152 in K562-IFN

```
nnATTGCACAATnn
```

```
nnATTGTACAATnn
```

Diagram showing the effect of SNPs on TF binding.
• High-throughput techniques can be used to identify long range regulatory sequences.
• Regulatory sequences can have several functions (Epromoters).
• Variation lying within Epromoters might be related to relevant human diseases.
• Variants within Epromoters can potentially affect regulatory interactions.
Regulatory Genomics Lab
Karen Nuñez
Ana Villaseñor
Luis Macías
Monica Padilla
Lucia Ramirez
Ana León
Mayra Padilla
Oscar Aldana

Collaborators
Morgane Thomas-Chollier (ENS-Paris)
Jacques van Helden (Aix Marseille)
Salvatore Spicuglia (Aix Marseille)
Jessica Dennis (UBC)
Nancy Cox (UV)

LIIGH
Jair García
Luís Aguilar

@AleMedinaRivera
amedina@liigh.unam.mx
http://www.liigh.unam.mx/amedina
Venue: UNAM-Campus Juriquilla

Dates: 25th - 30th October 2020

Special Issue in the Journal of the Royal Society Interface.

Call for papers opens February 2020
Confirmed Speakers:

- **Terry Gaasterland**, Scripps
- **Sandrine Dudoit**, UC Berkeley
- **Ludmil Alexandrov**, UCSD
- **Morgane Thomas-Chollier**, ENS-Paris
- **Dave Adams**, Wellcome Trust
- **Steve Hoffman**, Leibniz Institute for Aging
- **Katja Nowick**, Freie Universität Berlin

- **Manuel Corpas**, Cambridge Precision Medicine Ltd
- **Selene Fernández Valverde**, LANGEBIO
- **Peter Stadler**, Universidad de Leipzig, Max Planck Institute
- **David Sankoff**, University of Ottawa
- **Gabriel Moreno**, University of Waterloo
Día de Muertos
November 1\textsuperscript{st} and 2\textsuperscript{nd}
Acknowledgments:

Regulatory Genomics Lab
Karen Nuñez
Ana Villaseñor
Luis Macías
Monica Padilla
Lucia Ramirez
Ana León
Mayra Padilla
Oscar Aldana

Collaborators
Morgane Thomas-Chollier (ENS-Paris)
Jacques van Helden (Aix Marseille)
Salvatore Spicuglia (Aix Marseille)
Jessica Dennis (UBC)
Nancy Cox (UV)

LIIGH
Jair García
Luís Aguilar

http://www.liigh.unam.mx/amedina
@AleMedinaRivera
amedina@liigh.unam.mx
Which are the genes regulated by Epromoters?
Which are the genes regulated by Epromoters?

Lan Dao, et al, Nature Genetics. 2017
Which are the genes regulated by Epromoters?

ChIA-PET
P–P interaction

P1
P2

Lan Dao, et al, Nature Genetics. 2017
Which are the genes regulated by Epromoters?

Lan Dao, et al, Nature Genetics. 2017
Which are the genes regulated by Epromoters?

What is the effect of Epromoters on long range regulated genes?

Cell type A

Inverse correlation?

Positive correlation?
What is the effect of Epromoters on long range regulated genes?

Cell type A

Inverse correlation?

Positive correlation?

Cell type B
Genotype Tissue Expression Project

https://commonfund.nih.gov/gtex
eQTLs can have negative or positive correlation to gene expression

- Beta value is a measure of the effect the variant has on gene expression.
- Beta values is bound between -2 and 2.
Are eQTLs overlapping Epromoters?

Lan Dao, et al, Nature Genetics. 2017
It is more likely to find an eQTL in an Epromoter than in a non-Epromoter

\[ P = 7.28 \times 10^{-6} \]

Lan Dao, et al, Nature Genetics. 2017
What is the effect of Epromoters on long range regulated genes?

Lan Dao, et al, Nature Genetics. 2017
Epromoter eQTLs have a significantly stronger effect on distal gene expression

Lan Dao, et al, Nature Genetics. 2017
What is the effect of eQTLs that disrupt TF binding?

variation-scan

nnATTG
nnATTGT
eQTL disrupting TF binding tend to have a positive beta-value
Challenging biological concepts
The Non-Coding Genome

• Around 80% of the genome has shown to have some biochemical interaction.

• Most non-coding region have a regulatory function.

How is regulation being affected by genetic variants?
Promoter interactions detected with Pol II
ChIA-PET

Major challenges to connect the regulatory genome to human pathologies

- Detect long-range regulatory regions (enhancers) and connect them to the regulated genes.

- Evaluate the effect genetic variants within regulatory sequences have on gene expression.
Transcription factors control gene regulation by binding to specific DNA sequences.

- Transcription Factors interact with DNA binding to sequence specific sites.
Scanning a peak sequence with a TF binding motif

- Scanning

**Sox2**

...AGTGTGGACCATGAAATCTCTAACAGGTCTGATACTGAGGGA...

Matrix score

Sequence position (0 = center of the peak fragment)
Detecting protein disruption of binding at base pair resolution

- Genetic variants in gene regulatory regions are known to cause many diseases
- Can be informative of transcription factors and regulatory mechanisms
Detecting protein disruption of binding at base pair resolution

- Genetic variants in gene regulatory regions are known to cause many diseases
- Can be informative of transcription factors and regulatory mechanisms